Data Owner 1:

Walker 31.07 xxx
Sophia 11.04 xxx
Clerk 11.12 xxx
Jana 21.11 xxx

Bloom Filter based encoding

10101 xxx
11001 xxx
01101 xxx
10010 xxx

Exchange Encoding Secrets
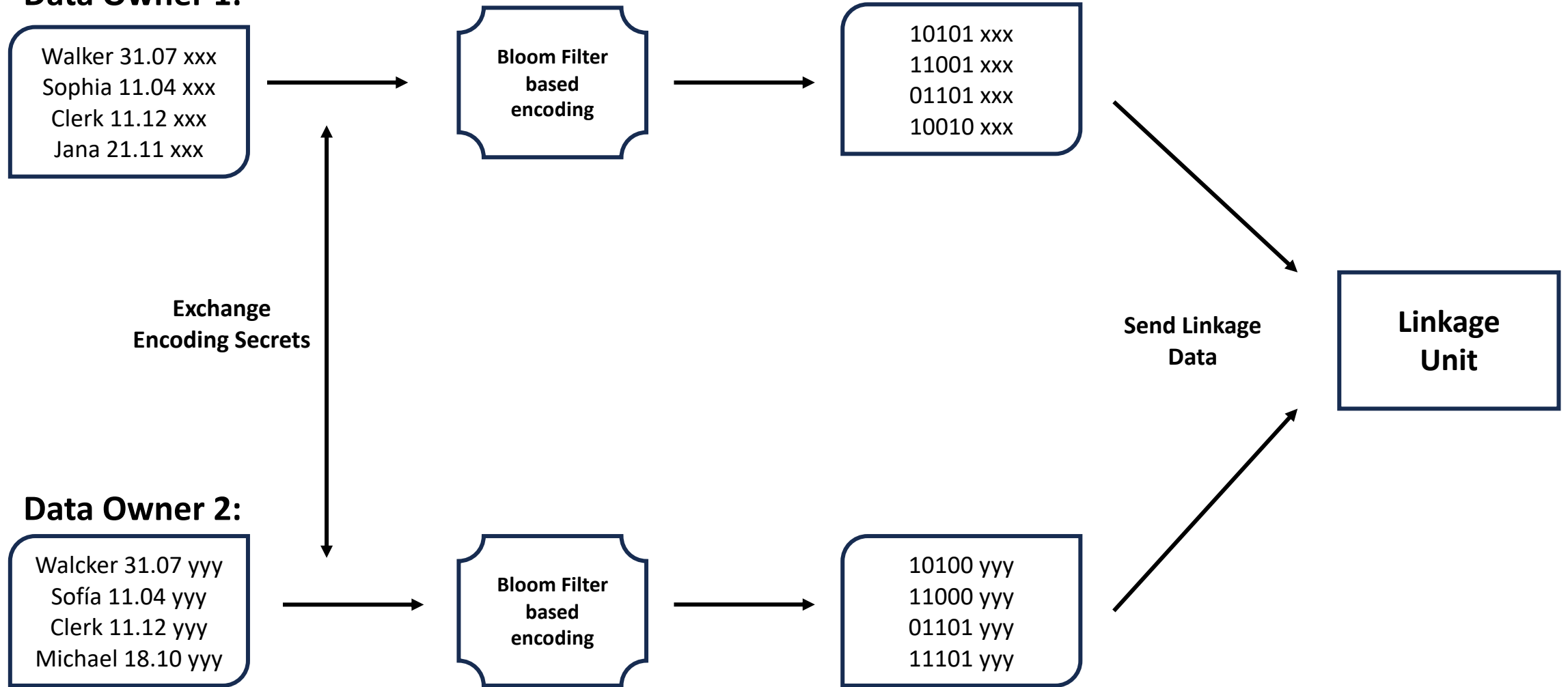
Data Owner 2:

Walcker 31.07 yyy
Sofía 11.04 yyy
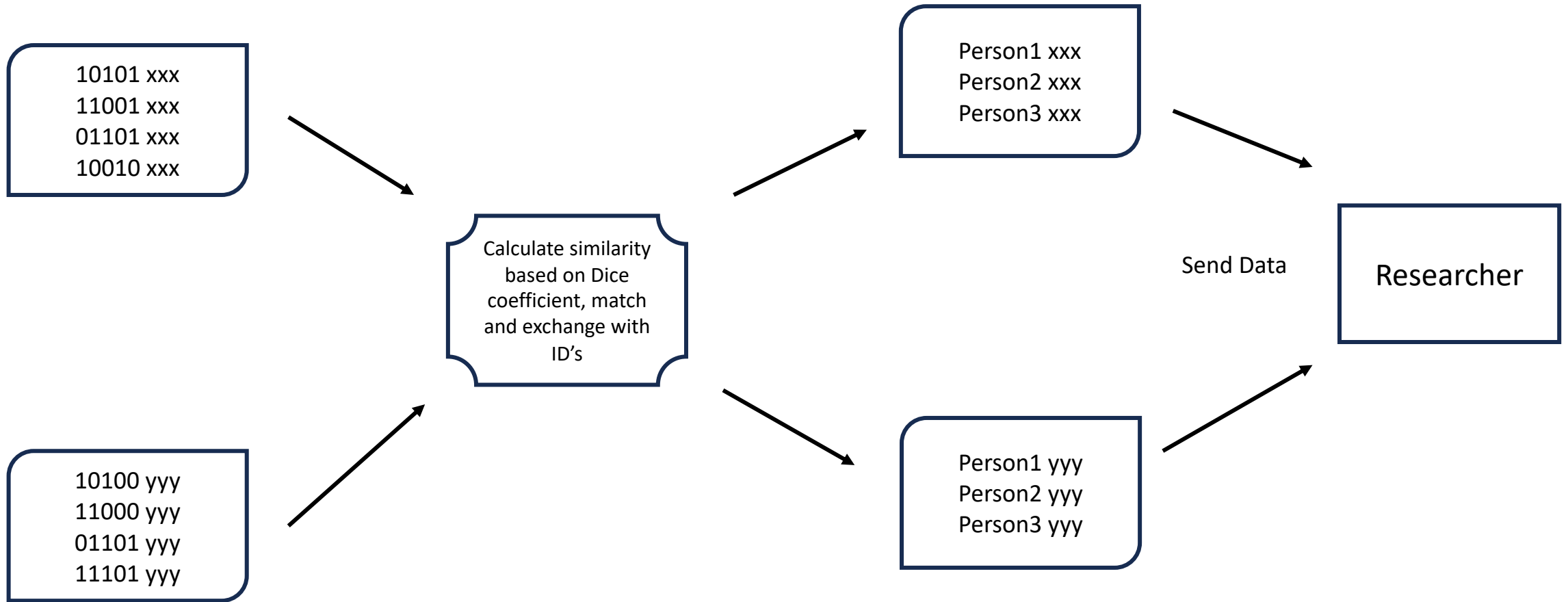Clerk 11.12 yyy
Michael 18.10 yyy

Bloom Filter based encoding

10100 yyy
11000 yyy
01101 yyy
11101 yyy

Send Linkage Data

Linkage Unit

Linkage Unit:

10101 xxx
11001 xxx
01101 xxx
10010 xxx

10100 yyy
11000 yyy
01101 yyy
11101 yyy

Calculate similarity based on Dice coefficient, match and exchange with ID's

Person1 xxx
Person2 xxx
Person3 xxx

Person1 yyy
Person2 yyy
Person3 yyy

Send Data

Researcher

Researcher:

Person1 xxx
Person2 xxx
Person3 xxx

Person1 yyy
Person2 yyy
Person3 yyy

Join
Data

Person1 xxx yyy
Person2 xxx yyy
Person3 xxx yyy

{ {en}, {nc}, {co}, {od}, {di}, {in}, {ng} }

| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
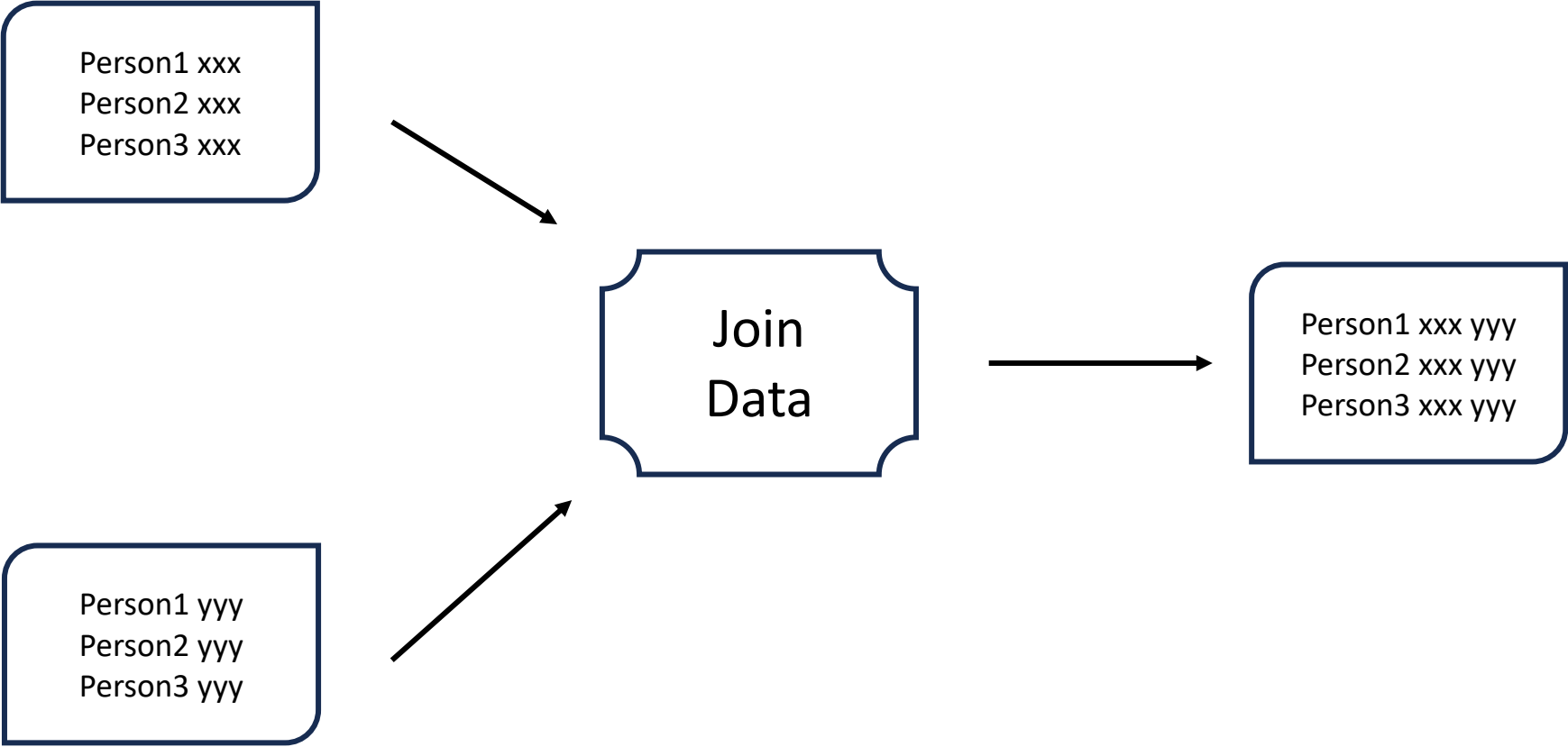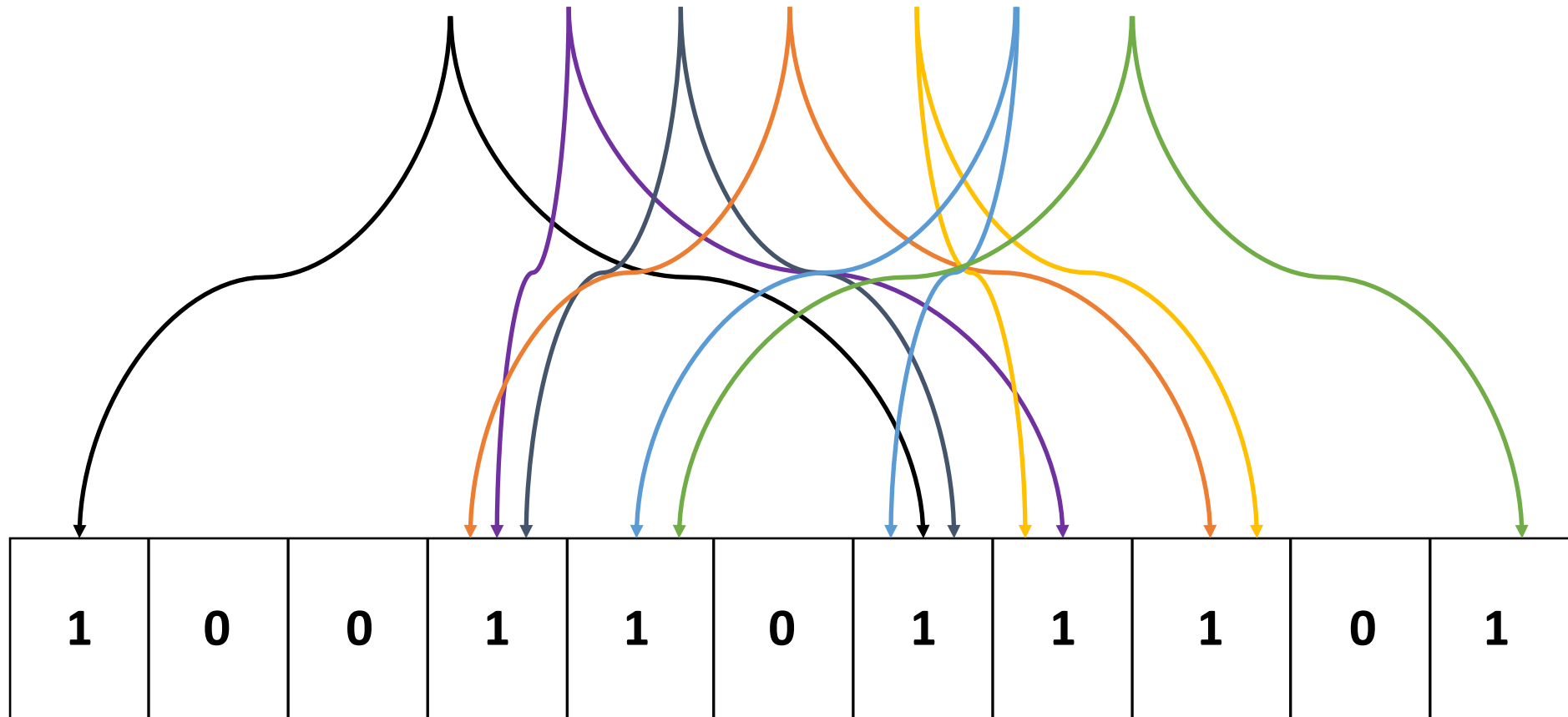
Index	Element	S	S'
1	1	1	1
2	2	0	1
3	3	1	1
4	4	1	0

π_1

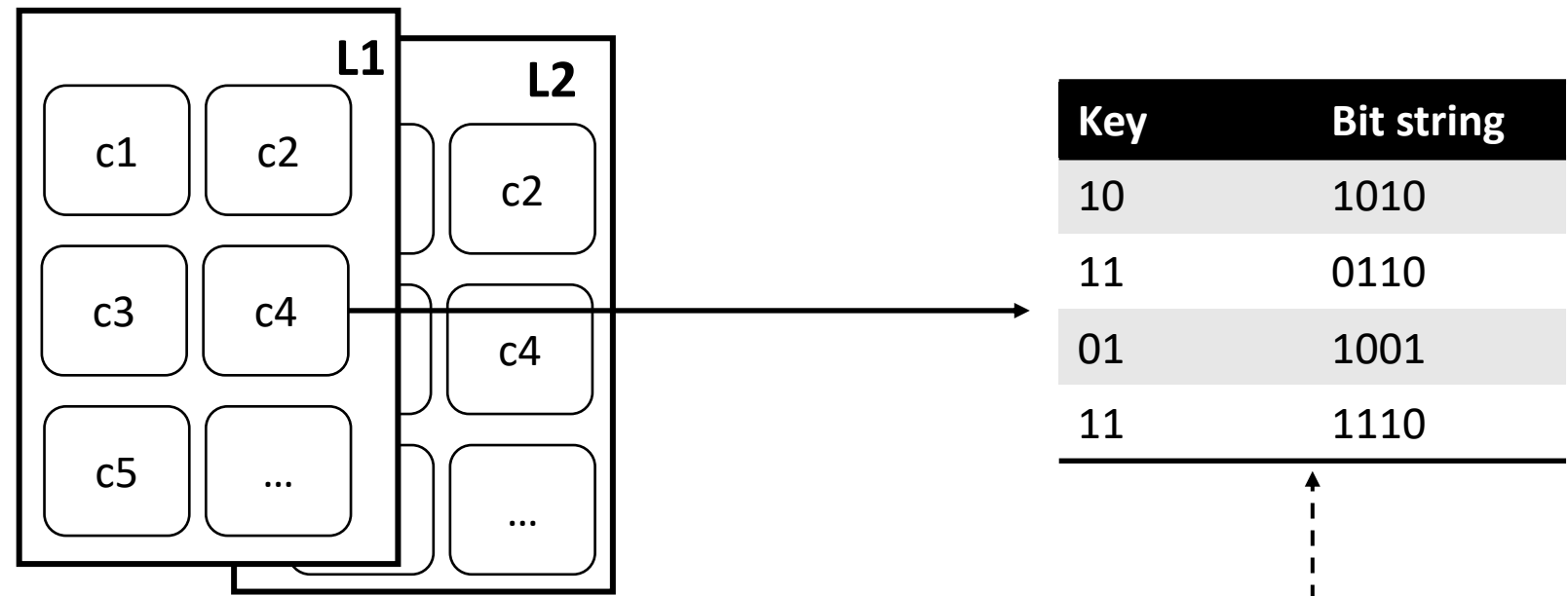| Index	Element	S	S'
1	4	1	0
2	3	1	1
3	2	0	1
4	1	1	1

Estimated Jaccard Similarity
$$= [\ 1(\pi_1(S_1)=\pi_1(S'_1)) + 1(\pi_2(S_1)=\pi_2(S'_1)) \ ] / 2$$
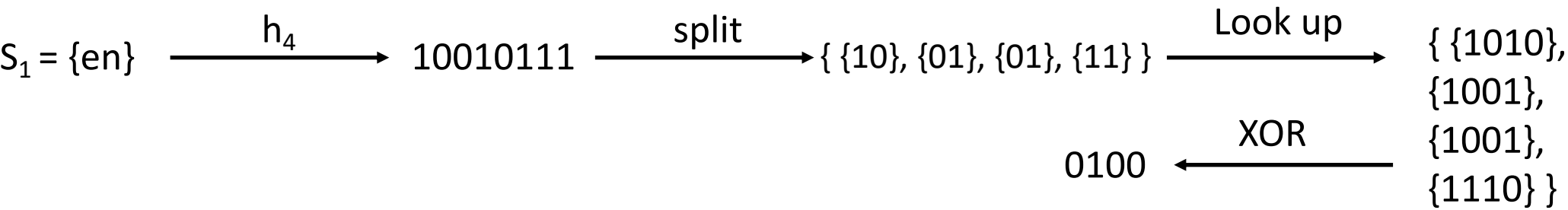$$= (1 * 0 + 1 * 1) / 2$$
$$= 1/2$$

π_2

Index	Element	S	S'
1	3	1	1
2	4	1	0
3	2	0	1
4	1	1	1

1) Initialization

L1

c1	c2
c3	c4
c5	...

L2

c2
c4
...

Key	Bit string
10	1010
11	0110
01	1001
11	1110

S = { {en}, {nc}, {co}, {od}, {di}, {in}, {ng} }

2) Hashing Process

S_1 = {en} $\xrightarrow{h_4}$ 10010111 $\xrightarrow{split}$ { {10}, {01}, {01}, {11} } $\xrightarrow{\text{Look up}}$ { {1010}, {1001}, {1001}, {1110} }

0100 $\xleftarrow{\text{XOR}}$

$\mathcal{Q}_1 = \{\text{pe, et, te, er}\}$

$\mathcal{H}_1$	**0**	**1**	0	**1**	**0**	**1**	1	**0**
$\mathcal{H}_2$	**1**	**1**	0	**0**	**1**	**0**	1	**0**
$\mathcal{H}_3$	**1**	**0**	0	**1**	**1**	**0**	1	**0**
$\mathcal{H}_4$	**0**	**0**	1	**1**	**0**	**1**	0	**1**

$\mathcal{G}(\mathcal{B}_\mathrm{P})$ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

$E_1 = \{\ \mathbf{53}\ ,\mathbf{113}\ ,\ 42\ ,\ \mathbf{7}\ ,\ \mathbf{256}\ ,\ \mathbf{87}\ ,\ 101\ ,\ \mathbf{21}\ \}$

$\mathcal{Q}_2 = \{\text{pe, et, te}\}$

0	**1**	0	**1**	**0**	**1**	0	**0**
1	**1**	0	**0**	**1**	**0**	0	**0**
1	**0**	0	**1**	**1**	**0**	0	**0**
0	**0**	0	**1**	**0**	**1**	0	**1**

↓ ↓ ↓ ↓ ↓ ↓

$E_2 = \{\ \mathbf{53}\ ,\mathbf{113}\ ,\quad\ \mathbf{7}\ ,\ \mathbf{256}\ ,\ \mathbf{87}\ ,\quad\ \mathbf{21}\ \}$

$Sim_\mathrm{J}(\mathcal{Q}_1, \mathcal{Q}_2) = \dfrac{|\mathcal{Q}_1 \cap \mathcal{Q}_2|}{|\mathcal{Q}_1 \cup \mathcal{Q}_2|}$

$= 0.75$

$Sim_\mathrm{J}(E_1, E_2) = \dfrac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$

$= 0.75$

S = { {en}, {nc}, {co}, {od}, {di}, {in}, {ng} } → **hash** →

	l1	l2	l3	l4
k_1	1	1	1	0
k_2	1	0	0	0
k_3	0	0	1	0
k_4	1	1	0	0

h'_1 ↓ h'_2 ↓ h'_3 ↓   **skip** ✖

22   **13**   **8**

S = { {22}, {13}, {8} } ← **result**

Data Owner 1:

Phillip 31.07 xxx
Sophia 11.04 xxx
Clerk 11.12 xxx
Jana 18.10 xxx

Bloom Filter based encoding

10101 xxx
11001 xxx
01101 xxx
10010 xxx

Exchange Encoding Secrets

Send Linkage Data

Linkage Unit

Data Owner 2:

Fillip 31.07 yyy
Sofía 11.04 yyy
Clerk 11.12 yyy
Michael 18.10 yyy

Bloom Filter based encoding
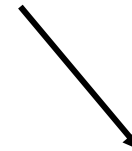
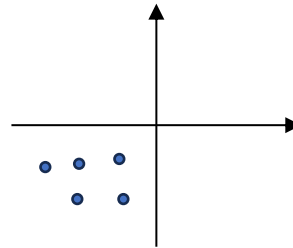10100 yyy
11000 yyy
01101 yyy
11101 yyy
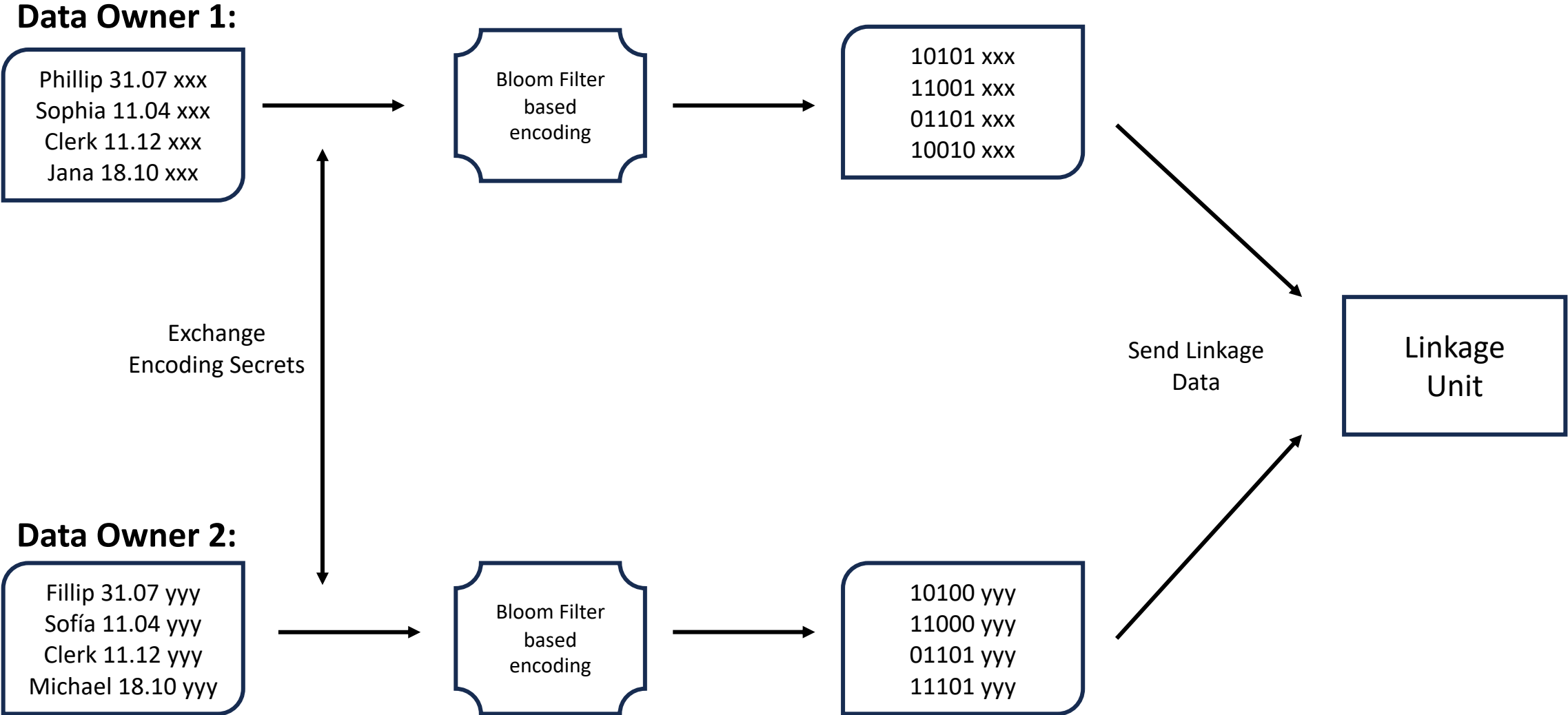
Linkage Unit:

10101 xxx
11001 xxx
01101 xxx
10010 xxx

Construct similarity graph and embeddings

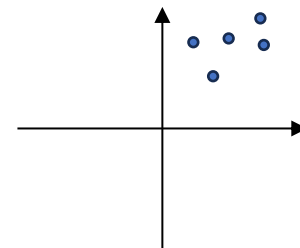Re-identify records by aligning embeddings and bipartite graph matching

Sophia 11.04 xxx
Clerk 11.12 xxx
Jana 21.11 xxx

Public Phonebook:

Sophia 11.04
Clerk 11.12
Jana 21.11
Michael 18.10

Mimic BF encoding and construct similarity graph with embeddings

Data Owner 1:

Walker 31.07 xxx
Sophia 11.04 xxx
Clerk 11.12 xxx
Jana 18.10 xxx

→ Bloom Filter based encoding →

10101 xxx
11001 xxx
01101 xxx
10010 xxx

Exchange Encoding Secrets

Data Owner 2:

Walcker 31.07 yyy
Sofía 11.04 yyy
Clerk 11.12 yyy
Michael 18.10 yyy

→ Bloom Filter based encoding →

10100 yyy
11000 yyy
01101 yyy
11101 yyy

Send Linkage Data

Linkage Unit

Linkage Unit:

10101 xxx 11001 xxx 01101 xxx 10010 xxx	→	Similarity Graph and Embeddings

Public Phonebook:

Sophia 11.04 Clerk 11.12 Jana 21.11 Michael 18.10	→	Bloom Filter based encoding (guessing), Similarity Graph and Embeddings

Align and re-identify intersection of data

Sophia 11.04 xxx
Clerk 11.12 xxx
Jana 21.11 xxx

Linkage Unit:

Known:

10101 **Sophia 11.04** xxx
11001 **Clerk 11.12** xxx
01101 **Jana 21.11** xxx

Train Neural
Network

Unknown:

10010 xxx
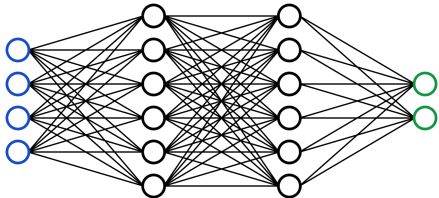
10010 **Walker 31.07** xxx

```
┌─────────────────────┐      ┌─────────────────────┐      ┌─────────────────────┐
│                     │      │                     │      │                     │
│                     │  ▶   │                     │  ▶   │   Hyperparameter    │
│ Graph Matching Attack│      │  Data Preparation   │      │    Optimization     │
│                     │      │                     │      │                     │
│                     │      │                     │      │                     │
└─────────────────────┘      └─────────────────────┘      └─────────────────────┘
                                                                     │
                                                                     ▼
┌─────────────────────┐      ┌─────────────────────┐      ┌─────────────────────┐
│                     │      │                     │      │                     │
│   Refinement and    │  ◀   │ Application to       │  ◀   │                     │
│   Reconstruction    │      │ Encoded Data         │      │   Model Training    │
│                     │      │                     │      │                     │
│                     │      │                     │      │                     │
└─────────────────────┘      └─────────────────────┘      └─────────────────────┘
```
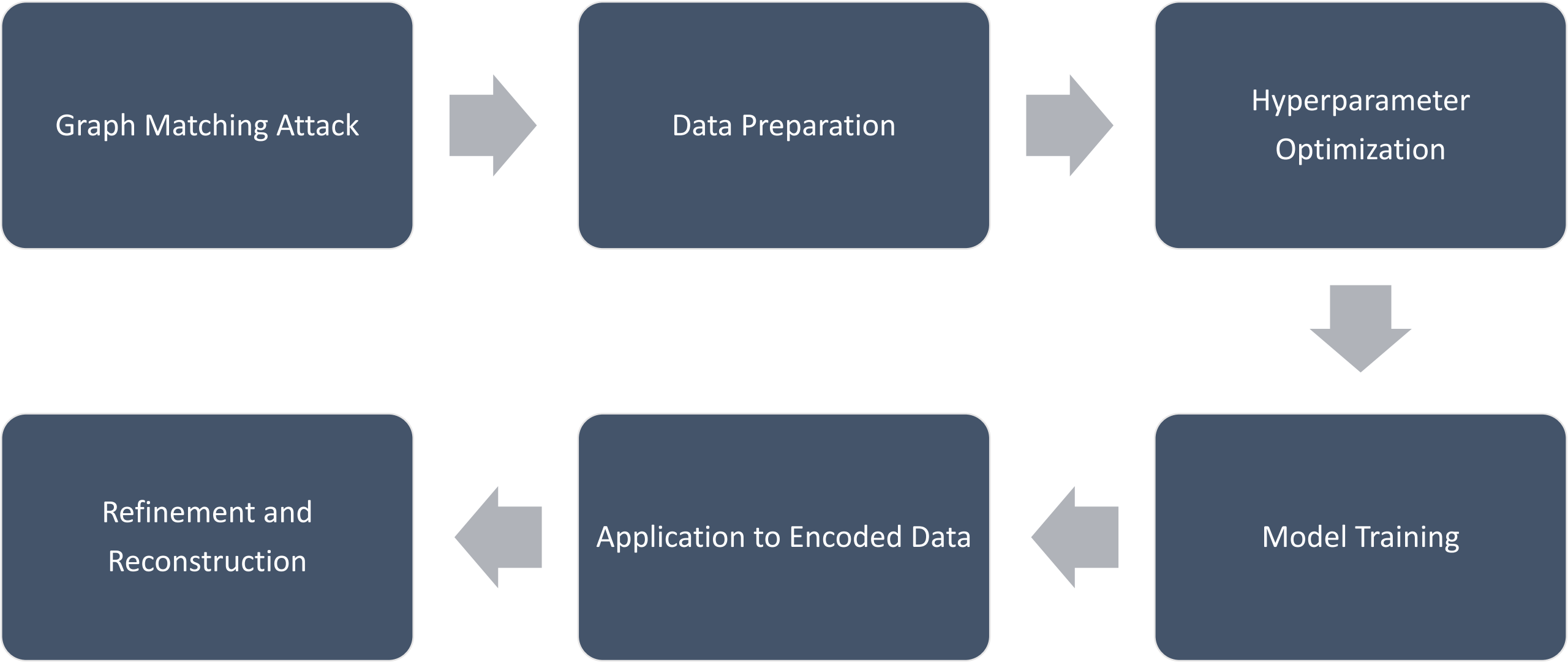
Input layer
i

Hidden layers
h_1 h_2 h_n

Output layer
o

Input 1

Input 2

Input n

Output 1

Output n